Analisi Predittiva

CT0429
Primo appello

Gennaio, 2022

Cognome: ______________________ Nome: ______________________

Matricola: ______________________ Firma: ______________________

Istruzioni
(da leggere attentamente).

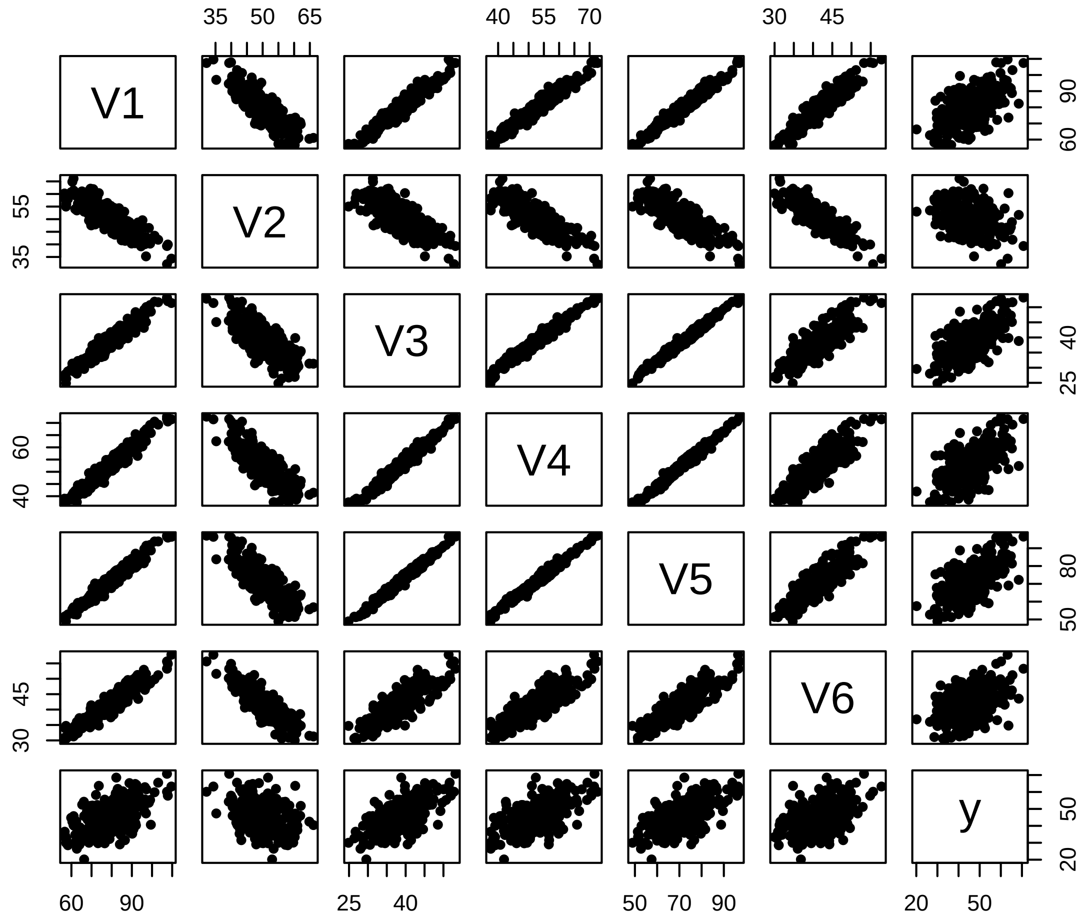
Assicuratevi di aver scritto nome cognome e matricola sia qui che sul file Rmarkdown disponibile su Moodle. Il tempo a disposizione per completare tutto l'esame (la parte scritta e la parte su Moodle) è di **90 minuti**.

Nessuno studente può lasciare l'aula fino a che la docente non avrà verificato che tutti abbiano consegnato sia il compito scritto che il file Rmarkdown. Dopo la consegna attendete che la docente dia il permesso di lasciare l'aula.

Question 1 (4 points)

Si prenda in considerazione il dataset `df`: i grafici di dispersione per le variabili nel dataset sono mostrati nella figura sottostante. Si desidera predirre la variabile `y` usando come predittori le variabili `V1, V2, V3, V4, V5, V6`.

```
plot(df, pch = 16)
```



Per costruire un modello di regressione multiplo un'analista utilizza in prima istanza un modello (chiamato `fitAll`) in cui tutti i predittori sono inseriti come variabili esplicative. Inoltre stima anche un modello `fitV4` in cui solo la variabile V4 viene usata come predittore. Informazioni riassuntive sulla stima dei due modelli sono mostrati nella pagina sucessiva.

```
 fitAll <- lm(y~., data = df)
 summary(fitAll)

Call:
lm(formula = y ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.1723  -4.0940   0.0372   4.1096  21.4763 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -50.8604    14.3392  -3.547 0.000454 ***
V1           -0.4226     0.3567  -1.185 0.236994    
V2            0.6375     0.1553   4.106 5.23e-05 ***
V3            0.2563     0.6587   0.389 0.697434    
V4           -0.3347     0.4459  -0.750 0.453561    
V5            1.3583     0.5456   2.489 0.013351 *  
V6            0.2098     0.3580   0.586 0.558244    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.211 on 293 degrees of freedom
Multiple R-squared:  0.4878,	Adjusted R-squared:  0.4773 
F-statistic:  46.5 on 6 and 293 DF,  p-value: < 2.2e-16

 fitV4 <- lm(y~V4, data = df)
 summary(fitV4)

Call:
lm(formula = y ~ V4, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.5657  -4.6950  -0.5004   4.7360  23.0933 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.72954    3.04571   1.881   0.0609 .  
V4           0.75756    0.05723  13.238   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.828 on 298 degrees of freedom
Multiple R-squared:  0.3703,	Adjusted R-squared:  0.3682 
F-statistic: 175.2 on 1 and 298 DF,  p-value: < 2.2e-16
```

(i) Qual è il valore di R^2 del modello per il modello `fitAll`? Come si può interpretare il valore di R^2 per questo modello?

R^2 per fitAll e': 0.4878

Interpretazione: "Circa il 48.78% della varianza osservata nei dati e' spiegata dalla relazione lineare tra Y ed i predittori V1, V2, V3, V4, V5, V6".

(ii) Come si può interpretare il valore del coefficiente angolare relativo alla variabile `V4` nei due modelli stimati?

In fitV4, V4 rappresenta il coefficiente angolare (in questo caso molto significativo) che detta il cambiamento stimato tra due istanze aventi un'unita' di V4 di differenza.
In poche parole, due osservazioni aventi un'unita' di misura di V4 di differenza, avrebbero una differenza di 0.75756 unita' di misura di Y.

In fitAll, V4 e' uno dei tanti predittori del modello. Il significato di V4 viene interpretato come il cambiamento della risposta, di -0.3347 unita' di misura di Y, tra due istanze che differiscono per un'unita' di misura di V4 (assumendo che il resto dei predittori abbiano dei valori fissati).

(iii) In calce sono indicati dei valori di Variance Inflation Factors (VIFs) per dei modelli stimati. Quale delle due opzioni è più probabile corrisponda ai VIFs per il modello `fitAll` stimato usando i dati mostrati nella Figura? Opzione ___2___

Opzione 1:

```
car::vif(fitAll)

       V1         V2         V3         V4         V5         V6
94.303945   5.439288  87.504215  73.386967 188.780574  22.216378
```

Opzione 2:

```
car::vif(fitAll)

      V1       V2       V3       V4       V5       V6
1.025082 1.018174 1.006091 1.046011 1.049240 1.022209
```

Question 2 (3 points)

Il gestore di un chiosco di gelati desidera studiare la relazione tra il numero di auto parcheggiate in una giornata nel parcheggio della spiaggia dove è posizionato il chiosco e il fatturato giornaliero (in decine di euro). Le informazioni disponibili al gestore sono salvate nel dataframe `df`. Le seguenti informazioni riassuntive sul dataset sono disponibili:

```
summary(df)
```

```
    numAuto         fatturato
 Min.   : 29.0   Min.   : 22.6
 1st Qu.:150.2   1st Qu.:112.8
 Median :217.5   Median :164.8
 Mean   :222.7   Mean   :166.7
 3rd Qu.:282.8   3rd Qu.:207.5
 Max.   :636.0   Max.   :423.5
```

Inoltre, il gestore stima il seguente modello predittivo per il fatturato:

```
fit <- lm(fatturato ~ numAuto, data = df)
```

```
summary(fit)
```

```
Call:
lm(formula = fatturato ~ numAuto, data = df)

Residuals:
   Min     1Q Median     3Q    Max
 -51.5  -11.8    0.4   13.1   38.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.810      5.954       3    0.004 **
numAuto        0.669      0.024      28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19 on 58 degrees of freedom
Multiple R-squared:  0.93,       Adjusted R-squared:  0.93
F-statistic: 7.6e+02 on 1 and 58 DF,  p-value: <2e-16
```

```
confint(fit)
```

```
                2.5 %     97.5 %
(Intercept) 5.8905551 29.7289948
numAuto     0.6202373  0.7176025
```

Il gestore desidera verificare l'evidenza contro l'ipotesi nulla che β_1 (il coefficiente angolare che descrive l'effetto di `numAuto` sul fatturato) abbia valore pari a 0.6, cioè vuole condurre un test di verifica di ipotesi per il sistema di ipotesi:

$$H_0 : \beta_1 = 0.6 \qquad VS \qquad H_1 : \beta_1 \neq 0.6$$

Il gestore desidera infine predirre il fatturato per una giornata in cui sonno presenti 300 auto nel parcheggio:

```
nd <- data.frame(numAuto = 300)
predict(fit, newdata = nd)
[1] NA
```

(i) Si derivi il valore della statistica test per il sistema di verifica di ipotesi specificato nel testo

(ii) Si indichi se è possibile o meno rigettare l'ipotesi nulla del sistema di verifica di ipotesi specificato nel testo (al livello di significatività del 5%)

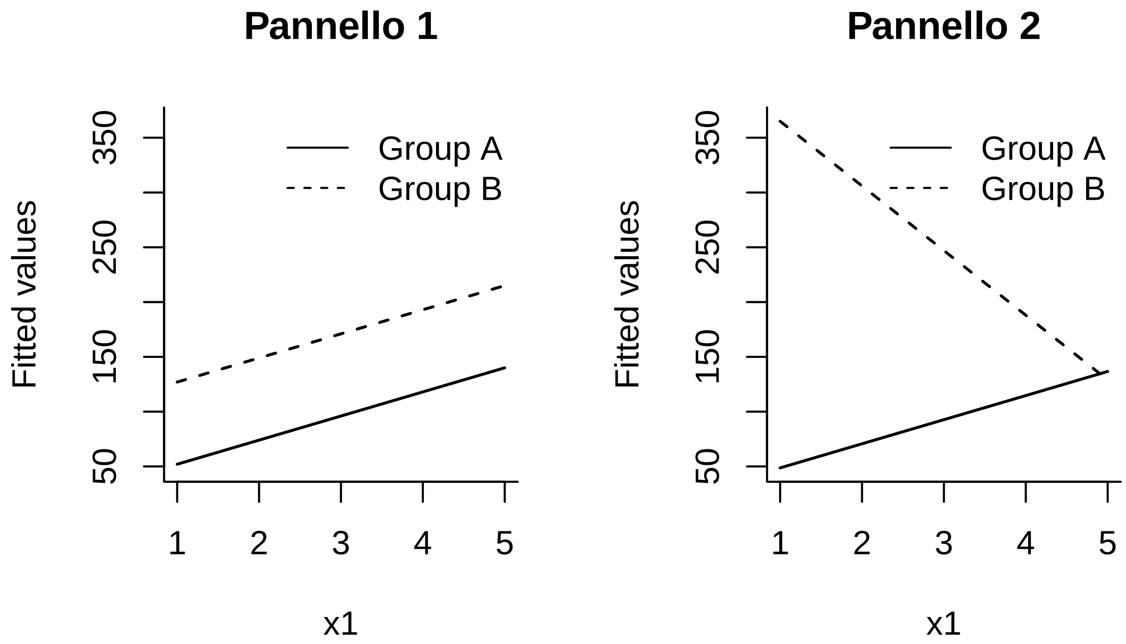
(iii) Si indichi il valore stimato del fatturato (in decine di euro) per una giornata in cui 300 auto sono presenti nel parcheggio (si indichi cioè il valore mancante dell'output di `predict`)

Question 3 (3 points)
Si prenda in considerazione il seguente modello stimato usando il dataset `df` mostrato nel codice sottostante:

```
fit <- lm(y ~ x1+x2+x1:x2, data = df); coef(fit)

(Intercept)          x1         x2b       x1:x2b
   26.66667    22.00000   397.33333   -81.00000

df

  x1 x2   y
1  1  a  45
2  2  a  78
3  3  a  89
4  4  b 188
5  5  b 129
```

(i) Si dia l'espressione della matrice di disegno X usata nella stima del modello (quello cioè che si otterrebbe usando `model.matrix(fit)`).

(ii) La Figura nella pagina successiva mostra la relazione stimata tra x1 e y per i due gruppi A e B (l'informazione del gruppo di appartenenza per ogni osservazione è specificata nella variabile x2). In quale dei due pannelli è più probabile che sia mostrata la relazione stimata dal modello `fit`)? Pannello ________


Pannello 1
Fitted values
50
150
250
350
Group A
Group B
1
2
3
4
5
x1
Pannello 2
Fitted values
50
150
250
350
Group A
Group B
1
2
3
4
5
x1

Question 4 (5 points)

Il gestore di un chiosco di gelati desidera studiare la relazione tra il numero di di gelati venduti tra le 14 e le 18 di una giornata e alcuni potenziali predittori: la temperatura della giornata, l'informazione se la giornata è un giorno festivo, il numero di auto presenti nel parcheggio della spiaggia dove è posizionato il chiosco.

Stima quindi tre diversi modelli per cui sono fornite alcune informazioni sintetiche:

```
fit1 <- glm(numGelati ~ numAuto+festivo, data = df, family = poisson)
fit2 <- glm(numGelati ~ numAuto+temperatura, data = df, family = poisson)
fitAll <- glm(numGelati ~ numAuto+temperatura+festivo,
              data = df, family = poisson)
```

```
deviance(fit1); deviance(fit2); deviance(fitAll)
```

```
[1] 293.5679

[1] 94.42806

[1] 68.08272
```

```
logLik(fit1); logLik(fit2); logLik(fitAll)
```

```
'log Lik.' -217.192 (df=3)

'log Lik.' -117.6221 (df=3)

'log Lik.' -104.4494 (df=4)
```

```
coef(fit1)
```

```
  (Intercept)        numAuto festivofestivo
        1.50           2.00           0.24
```

```
nd <- data.frame(numAuto = 100, festivo  = "feriale", temperatura = 25)
```

```
predict(fit, newdata = nd, type = "response")
```

```
[1] NA
```

(i) Usando il modello `fit1` si stimi il numero di gelati venduti in una giornata feriale in cui 100 auto sono presenti nel parcheggio e la temperatura è di 25 gradi (si indichi cioè il valore mancante dell'output di `predict`)

(ii) Quale modello tra i tre stimati raggiunge un valore di AIC minore?

(iii) Come è possibile confrontare la bontà di adattamento dei modelli `fit1` e `fit2`? E se invece si desidera confrontare la bontà di adattamento dei modelli `fit1` e `fitAll`? (non è necessario confrontare effettivamente i modelli, basta descrivere come sarebbe possibile confrontarli).

Question 5 (3 points)

In un sito di e-commerce viene monitorato se per una determinata transazione viene esercitata l'opzione di reso: questa informazione è salvata nella variabile `reso`, che ha valore 1 se per la transazione è stata esercitata l'opzione di reso. Un primo modello predittivo mira a verificare se l'ammontare totale del costo della transazione (variabile `totalSpent`) influisce sulla probabilità che venga attivata l'opzione di reso:

```
fit1 <- glm(reso ~ totalSpent, data = df, family = binomial())
coef(fit1)
```

```
(Intercept)  totalSpent
-7.60140758  0.03365754
```

Viene poi stimato un altro modello in cui come predittori vengono anche inserite delle variabili che indicano il numero totale di pezzi nell'ordine (`totalPieces`) e l'informazione se la consegna dell'ordine è avvenuta in ritardo (`isDelayed`):

```
fit2 <- glm(reso ~ totalSpent+totalPieces+isDelayed,
            data = df, family = binomial())
```

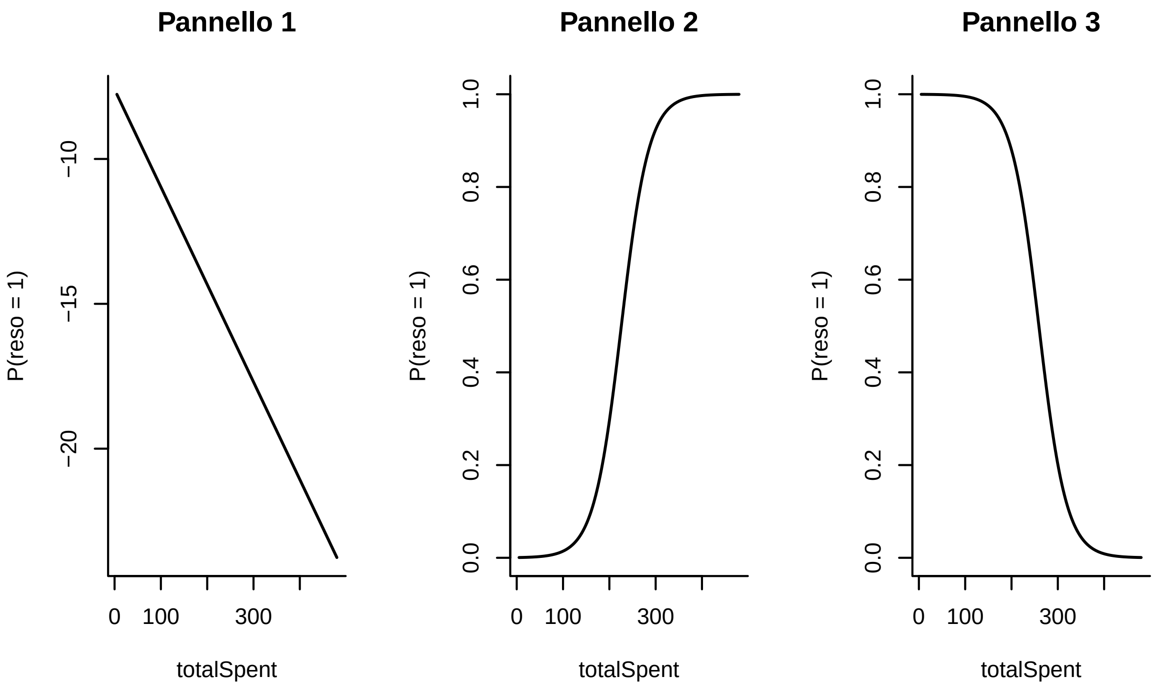
I due modelli vengono poi confrontati tramite un test `anova`:

```
anova(fit1, fit2, test = "LRT")
```

```
Analysis of Deviance Table

Model 1: reso ~ totalSpent
Model 2: reso ~ totalSpent + totalPieces + isDelayed
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       148     66.619
2       146     65.354  2   1.2651   0.5312
```

(i) La Figura nella pagina successiva mostra la relazione stimata tra `totalSpent` e la probabilità che per la transazione venga attivato un reso (cioè P(`reso` $= 1$)). In quale dei tre pannelli è più probabile che sia mostrata la relazione stimata dal modello `fit1`? Pannello ________

(ii) Cosa possiamo evincere dal test `anova` in cui vengono messi a confronto i modelli `fit1` e `fit2`?


Pannello 1
P(reso = 1)
−10
−15
−20
0 100 300
totalSpent
Pannello 2
P(reso = 1)
1.0
0.8
0.6
0.4
0.2
0.0
0 100 300
totalSpent
Pannello 3
P(reso = 1)
1.0
0.8
0.6
0.4
0.2
0.0
0 100 300
totalSpent